Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Member Name: • Shaik Ahmad Basha ahmadshaik982basha@gmail.com Contribution: • Exploring Data • Data Wrangling • Data Cleaning • Checking for Null Values and Duplicated Values • Analyze How Numerical Features Varies with Dependent Feature • Analyze How Categorical Features Varies with Dependent Feature • Analyze How Numerical Features Varies with Categorical Features • Performed EDA on Dataset • Identifying and Removing Outliers • Removing Multicollinearity • Feature Scaling • Handling Class Imbalance • Fitting the Data into Various Classification Models • Plotting Feature Importance. • Tuning the Hyperparameters to Avoid Overfitting
Please paste the GitHub Repo link.
GitHub Link: - https://github.com/ahmedshaik982/Cardiovascular_Risk_Prediction
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Problem Statement:

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

Approaches:

The first step imported all the necessary libraries like NumPy, Pandas etc. and then collected the data. I started with understanding the data like what are the columns and their meanings and data types.

After that, the second step is data preprocessing. Data preprocessing is a process where raw data is converted into clean data. The dataset has null values. I just replaced those with median and mode for numerical features and categorical features respectively. The Third step is to analyze the data using Exploratory data analysis techniques.

And then I moved on to Feature engineering and preprocessing where I removed outliers present in the data and I also removed multicollinearity. I also scaled the input features by using StandardScaler.

And the last step is fitting the training dataset into various classification models and evaluating the models using test dataset and plotting feature importance.

Conclusions:

By fitting the data into various classification models and evaluating with test data, we can conclude that,

For our problem statement, we have two cases
Case 1 : if a person has a disease but the model shows the person has no risk. For this case recall is the best evaluation metric.
Case 2 : if a person has no risk of disease but the models predicts the person has a risk. For this case, precision is the best evaluation metric.
When both of the above cases are important, we can use F1_score evaluation metric

If Recall is important, i.e., if we want to focus on case 1, then Logistic Regression and KNN models are preferrable.

If Precision is important i.e., if we want to focus on case 2, then Naive Bayes model is preferrable.

If F_Score is important, i.e., if we want to focus on both case 1 and case 2, then Logistic Regression model is preferrable.
For all the classification models, 'age' feature has more importance.